METHOD AND APPARATUS FOR CHARACTERIZING DOCUMENTS BASED ON CLUSTERS OF RELATED WORDS

ABSTRACT

One embodiment of the present invention provides a system characterizes a document with respect to clusters of conceptually related words. Upon receiving a document containing a set of words, the system selects "candidate clusters" of conceptually related words that are related to the set of words. These candidate clusters are selected using a model that explains how sets of words are generated from clusters of conceptually related words. Next, the system constructs a set of components to characterize the document, wherein the set of components includes components for candidate clusters. Each component in the set of components indicates a degree to which a corresponding candidate cluster is related to the set of words.